DOCUMENT RESUME

ED 462 434 TM 033 698

AUTHOR Buckendahl, Chad W.; Impara, James C.; Plake, Barbara S.
TITLE A Strategy for Evaluating District Developed Assessments for State Accountability.
PUB DATE 2001-10-00
NOTE 19p.; Paper presented at the Annual Meeting of the Mid-Western Educational Research Association (Chicago, IL, October 24-27, 2001).
PUB TYPE Reports - Descriptive (141) -- Speeches/Meeting Papers (150)
EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS *Accountability; Achievement Tests; Criterion Referenced Tests; Elementary Secondary Education; *Evaluation Methods; School Districts; State Programs; State Standards; Student Evaluation; *Test Construction; *Testing Programs

ABSTRACT

One Midwestern state has chosen a model of state assessment in which local school districts are responsible for developing the strategies to measure and report their students' performance on state-adopted content standards. When students are not measured on common instruments, district accountability becomes an added challenge. This paper presents a strategy for evaluating locally developed assessments as part of the state assessment system that can be used to inform the need for state accountability. An application of the strategy is described in which a 16-member District Assessment Evaluation Team was recruited to evaluate district assessment portfolios for the state. An analysis of inter-rater agreement on the common districts evaluated by all 19 raters (16 team members and 3 anchor raters) was performed. A goal of 80% agreement or higher was established, but there was some variance in the levels of agreement for each criterion and the overall ratings. Results indicate there was reasonable consistency among raters. Benefits of the proposed strategy include an emphasis on formative rather than summative feedback and on improving assessment strategies at the local level. Locally developed assessments that are aligned to the state's content standards and integrated into the district's curriculum are likely to produce data that is meaningful to the state, yet can inform instruction in the classroom. appended are tables of data used in the study. (Author/SLD)

Reproductions supplied by EDRS are the best that can be made
from the original document.

ED 462 434

A strategy for evaluating district developed assessments for state accountability.

PERMISSION TO REPRODUCE AND DISSEMINATE THIS MATERIAL HAS BEEN GRANTED BY

C. Buckendahl

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

1

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

☑ This document has been reproduced as received from the person or organization originating it.

☐ Minor changes have been made to improve reproduction quality.

• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

Chad W. Buckendahl
University of Nebraska – Lincoln

James C. Impara
University of Nebraska – Lincoln

Barbara S. Plake
University of Nebraska – Lincoln

Paper presented at the annual meeting of the

Midwestern Educational Research Association, Chicago, IL

TM033698

BEST COPY AVAILABLE

October, 2001

Abstract

Most states use a statewide assessment strategy to evaluate districts on common measures. One Midwestern state has chosen a different model of state assessment where local school districts are responsible for developing the strategies to measure and report their students performance on state adopted content standards. When students are not measured on common instruments, district accountability becomes an added challenge. This paper presents a strategy for evaluating locally developed assessments as part of the state assessment system that can be used to inform the need for state accountability. An application of the strategy is also included. Benefits of the proposed strategy include an emphasis on a formative rather than summative feedback and on improving assessment strategies at the local level. Locally developed assessments that are aligned to the state's content standards and are integrated into the district's curriculum will likely produce data that may be meaningful to the state, yet informs instruction in the classroom.

A strategy for evaluating district developed assessments for state accountability

Statewide assessment and accountability systems are common topics within the educational community. In response to external pressures, control over methods of accountability has shifted in many instances from the local jurisdiction (school districts) to the state jurisdiction (departments of education and legislative agencies). The shift in control is explained in part, by the increased attention popular media has given to the perceived shortcomings of public education and the need for accountability. A key component of President Bush's education plan (Bush, 2001), includes testing students across a number of grade levels and using their performance to make decisions that reward or sanction school districts.

Many state accountability models include similar provisions for performance on state mandated tests (e.g., Florida, North Carolina, Texas). However, more recent work in the area of assessment has sought to re-conceptualize the role of assessment in school districts and the classroom as going beyond the narrowly constructed state assessments (Diaz, 2001). More importantly, the role of curriculum, instruction, and assessment in informing learning is being examined because these are the linkages that state assessments have a difficult time demonstrating (Shepard, 2000). The appropriate relationship among these three areas is a balance that statewide assessments have yet to fully address.

There are a variety of accountability systems employed across the country. States that use a common assessment system as the primary component of their accountability systems have rank ordered school districts based on performance at individual grades and content area sub-tests (e.g., Georgia). Other states (e.g., Kentucky)

have rank ordered or rated districts on a composite index of district performance that considers both achievement measures and non-cognitive indicators such as socioeconomic status, limited English proficiency, or mobility. And still other states rank order or rate school or district performance using scale scores that are based on test performance on content specific instruments and do not include non-cognitive indicators (e.g., Maryland).

Overview of an "Uncommon" Assessment System

For states that do not have a common assessment system (one that is the same across all districts in the state), refined comparisons across school districts are suspect. The challenge to these "uncommon" assessment systems is to employ a strategy that downplays comparisons and focuses on formative evaluation. Such a system suggests rating rather than rank ordering district performance to reduce the narrower comparisons that are evident in many state accountability systems. Currently one state has chosen an "uncommon" model for their state assessment system. The foundation of this state's approach is at the local level where districts have the primary responsibility for determining strategies that measure student performance on state adopted content standards in reading/writing, mathematics, science, and social studies. Using a combination of measurement strategies, the districts develop individual assessment plans to measure the content standards. Each district's assessment plan may be unique except for the state's writing assessment that is administered across all districts. Content areas are phased in annually beginning with reading/writing.

These assessment plans are submitted to the state's department of education and reviewed prior to implementation. After the district's strategies are employed during the

academic year, information about the quality of the assessments and the students performance on those assessments are reported to the department of education. This information is used separately by the department of education to produce a state report card on the performance of school districts on the state's content standards. A proposed strategy that was used to evaluate the information submitted on the quality of the assessments is at the heart of this paper.

Strategy for Evaluating the Quality of District Assessments

Unlike states that use a common assessment system, the assessment systems employed across districts in this state may not be similar. As a result, an evaluation of the quality of those assessment systems judged on a common rubric is needed to better understand the subsequent levels of student performance on those assessments. The technical quality rating serves as an "equating" factor for performance because districts that have high levels of student performance and a high quality assessment system are perceived as more credible than districts that have high levels of student performance and a low quality assessment system. This technical quality component of the state's accountability system represents a new contribution to accountability research that has not been seen in states with common assessment systems.

The general procedures for employing this evaluation strategy begins with districts providing documentation (called assessment portfolios in this state) to the department of education that describes their overall assessment plan and contains information about the technical quality of the assessment strategies for measuring student performance on the content standards. Next, an external evaluation team is recruited and

trained on six criteria that will be used to evaluate the quality of the district's assessments. After training and calibration activities, the evaluation team members are sent an equivalent number of district portfolios on which they will conduct independent evaluations of quality relative to the six quality criteria. Included with the set of unique assessment portfolios are two common portfolios (unknown by the evaluators) to measure the level of inter-rater reliability and fairness among the raters.

When the evaluation team members complete their ratings of the assessment portfolios, they return both their review forms and the portfolios to the organizing agency. At this point, districts' ratings for each of the six criteria and the overall evaluation for each grade level are compiled in a database for reporting purposes. Last, the individual district evaluation forms with feedback are returned to the districts for use in subsequent assessment development and revision. Results from the assessment technical quality ratings are then included in a state report card that is disseminated statewide in the fall.

Methods and Procedures

A sixteen member District Assessment Evaluation Team (DAET) was recruited to evaluate district assessment portfolios for the state. All members of the team had experience in measurement and were broadly selected. The DAET included members from the following states: Rhode Island, Tennessee, Illinois, Michigan, Wisconsin, Iowa, Nebraska, Oklahoma, Texas, and California. DAET members had extensive experience ranging from developing or overseeing test development in local school districts to developing credentialing examinations. The DAET convened in a centralized location

during the second week of May, 2001, for a three-day training workshop. At the workshop the DAET was trained on a technical quality scoring rubric developed by the Buros Center for Testing (Plake & Impara, 2000) specifying the characteristics necessary to achieve a given rating on the technical quality criteria. The six technical quality criteria are as follows: a) alignment of the assessment to the content standards, b) students are given the opportunity to learn the material prior to assessment, c) assessments are free from bias or offensive language, d) assessments are developmentally appropriate, e) there is consistency in scoring, and f) mastery levels are appropriate. Approximately one day of the training was spent familiarizing the DAET with the state's assessment model and the requirements for each of the six quality criteria.

Beginning on the second day of the training workshop and continuing through the third day, the DAET examined a sample district assessment portfolio working in small groups to evaluate the quality of the process the district used for each criterion. After the small groups rated each criterion, the entire group reconvened to discuss their ratings and the rationale for the their ratings. This process was repeated for a second sample district portfolio with the DAET members individually rating the quality of each criterion. The rating scale that the DAET used to individually evaluate the six criteria was as follows: a) Met – no additional comments needed, b) Met – with additional comments needed, c) Met – Needs Improvement, and d) Not Met. Any ratings of Met-Needs Improvement or Not Met were accompanied by feedback and suggestions about how the district could improve their local processes to meet the expectations of the criterion. In addition, because the intent of the evaluation process was to provide formative feedback to the

school districts, DAET members were encouraged to provide comments on any criterion that could be improved.

In terms of the overall district classification, only two distinctions were made on a district's performance on a given criterion, met or not met. Met was defined as being met with or without comments or met – needs improvement. This definition was true for five of the six criteria. For one criterion (consistency in scoring) a more refined definition was needed. Because consistency in scoring or reliability was defined using a numerical characteristic of scores, threshold values were required for a district to receive a "fully" met versus a "met – needs improvement" rating. The threshold value to be "fully" met was set a .70 for objectively scored instruments and at 70% inter-rater agreement for subjectively scored instruments. This threshold is consistent with generally accepted measurement reliability values for making group decisions. Values of .50 and 50% for objectively and subjectively scored tests respectively, were set for a met – needs improvement rating. This was the only criterion on which there was a distinction between the ratings relative to the overall classification decision. Although districts could use a variety of strategies to meet this criterion, specific minimum values were defined to ensure the credibility of the assessment process and results.

After districts submitted information about the quality of their district assessments in late June, 2001, the DAET members were sent an equivalent number of district's assessment portfolios (22-27) to independently evaluate. To ensure an appropriate level of inter-rater reliability, two district's assessment portfolios were blindly sent to all raters to estimate the level of agreement. To establish the anchor

ratings on the six criteria and the overall classification for these common districts, the authors rated them in advance of the DAET. The overall classification is determined by the combination of ratings from the six quality criteria. All criteria are not equally weighted in this decision process. A matrix that shows how these overall ratings are determined is provided as Appendix A. DAET members were sent a representative sample of district portfolios stratified on size and geographic location in early July and were given five weeks to complete their reviews. The next section presents the results of an analysis of the inter-rater agreement on the common districts that were evaluated by all raters.

Results

Analyses for inter-rater agreement were conducted for the two common district assessment portfolios across the six quality criteria, the overall rating, and comments provided as feedback. Table 1 shows the percent agreement among the 19 total raters (16 DAET members and 3 "anchor" raters) for each of the six quality criteria and the overall rating. Agreement was defined as the percent of raters agreeing that a given criterion was Met or Not Met relative to the anchor ratings. Again, "Met" was defined as being met without comments, met with comments, or met - needs improvement.

[Insert Table 1 Here]

Although a goal of 80% agreement or higher was established, there was some variance in the levels of agreement for each criterion and the overall ratings. For district A, the raters had high levels of agreement for the first four criteria (alignment to standards, opportunity to learn, bias review, and developmental appropriateness), but lower levels

of agreement on the last two criteria (consistency in scoring and appropriate mastery levels). The agreement on the overall rating was much lower than desired. For District B, the grade levels employed different strategies to meet the criteria. Therefore, the raters were required to provide separate ratings on the criteria for each of the grades. Although there was generally higher levels of agreement on the more technical aspects (criteria 5 and 6) and the overall rating, there was lower agreement on criteria 2 and 3 (opportunity to learn and bias review).

Because a goal of this process was to provide formative feedback to school districts, it was also important to have consistency in the comments the raters provided. Thus, a second analysis was conducted to determine the level of agreement among the raters. This analysis focused on the feedback comments that were provided in the district review forms. Table 2 shows the breakdown of comments by criterion for District A.

[Insert Table 2 Here]

Table 2 above shows the percent of raters who wrote specific comments for feedback to District A on the quality of each criterion in their assessment portfolio. Although some reviewers provided more feedback than others, there was evidence of reasonable consistency in these comments. The reviewers' comments focused on requests for additional information about the process or procedures the district used to determine whether it met the quality criterion. Table 3 below shows this information for District B.

[Insert Table 3 Here]

Table 3 above shows the percent of raters who wrote specific comments for

feedback to District B on the quality of each criterion in their assessment portfolio. There was also evidence of reasonable consistency in the comments for this district. Most reviewers' comments focused on requests for additional information about the process or procedures the district used to determine whether it met the quality criterion. Again, this analysis of the feedback comments was conducted because the intent of the review process is to provide formative evaluation information. Thus, it was important to determine whether the raters were consistently providing appropriate feedback to the districts for them to use in their future assessment plans.

Discussion

For a state assessment system that does not rely on a single assessment strategy, there is a need for a mechanism that can measure all districts against common criteria if the accountability system seeks credibility from the broad spectrum of stakeholders. In an effort to balance considerations for both local control and accountability, one state has selected a rating of assessment technical quality as this mechanism. The rationale for including a rating of the technical quality of district assessments in the accountability system is that it is necessary for districts to demonstrate the psychometric soundness of the methods they are using to determine student performance. If districts were only asked to provide student performance estimates from their local assessments without evidence of the quality of the assessment strategies they are using to measure performance, it would raise concerns about interpreting the performance.

The technical quality of district assessment components represents characteristics of sound measurement practices that are applicable within in a district setting.

Psychometric characteristics of alignment (including both content and cognitive validity), opportunity to learn, freedom from bias, development appropriateness, consistency in scoring, and appropriateness of mastery levels provide evidence of the technical quality of districts' assessments and adds to the trustworthiness of the reported student performance.

The strategy that was used to evaluate the technical quality of districts' assessments was untested until recently and has limitations that need to be addressed. The low to moderate levels of agreement on the individual criterion ratings as well as the overall ratings are problematic. If these ratings are considered to be an integral part of the state's accountability model, there must be higher levels of agreement among the raters on these elements. These low levels of agreement suggest that the training activity was insufficient to calibrate the raters to a common understanding of the scoring rubric and overall rating matrix.

This low agreement among reviewers may be explained in part by the time lapse between when the raters were trained (May) and when they actually received the materials to rate (July). Another explanation is that the conception of the rubric was not consistent across the raters, specifically with regard to the more technical components of reliability and mastery levels. Because the overall rating is heavily influenced by these technical components, it is essential that the DAET have a common understanding of how to evaluate the relevant district assessment information related to these components. To improve the rater's agreement, we would suggest extensive training with examples of district assessment portfolios with a variety of characteristics to show the range of

materials the DAET would be reviewing. We would also suggest that this training occur closer to the time the materials would actually be sent to the evaluators to reduce the time lapse between training and operational reviewing.

This paper presented a strategy for evaluating locally developed assessments as part of a state assessment system that informs the need for state accountability. An illustration of how this strategy was employed was also included. A benefit of the proposed strategy is that it focuses on formative as opposed to summative feedback that encourages local control of the assessment strategies to measure district students. Locally developed assessments that are aligned to the state content standards and are integrated into district curriculum will likely produce data that may be meaningful to the state, yet informs instruction in the classroom. By considering the technical quality of districts' assessments as part of the state accountability system, common criteria under which all districts are rated is added to the system allowing for limited comparisons across districts. This new area in accountability research is encouraging because it provides some evidence that it may be possible to inform state needs without jeopardizing the utility of the information at the local level.

References

Bush, G.W. (2001). *No child left behind.* The plan can be accessed at: www.ed.gov/inits/nclb.

Diaz, M.E. (2001). Will reform based on standards and assessment make a difference in the 21st century? *Mid-Western Educational Researcher, 14*(1), 22-27.

Shepard, L.A. (2000). The role of assessment in a learning culture. *Educational Researcher, 29*(7), 4-14.

Plake, B.S. & Impara, J.C. (2000). *Technical Quality Criteria for Nebraska's District Assessments*. Lincoln, NE: Buros Center for Testing.

Appendix A

Technical Quality Overall Rating Matrix

	Exemplary	Very Good	Good	Acceptable	Unacceptable
Alignment to Standards	Met	Met	Met	Met	Not Met or
Opportunity to Learn	Met	Met	Met	Met	Not Met
Freedom from bias or offensive situations	Met	Met	Met or	Any rating	Any rating
Developmentally Appropriate	Met	Met	Met	Any rating	Any rating
Consistency in scoring	Met	Met or	Met-NI and	Any rating	Any rating
Mastery levels are appropriate	Met	Met	Not Met	Any rating	Any rating

Table 1.

Percent agreement among raters (n=19) for two common school districts[1].

Criterion	District A	District B-4th	District B-8th	District B–11th
1	100%	79%	84%	95%
2	100%	58%	58%	89%
3	100%	58%	58%	53%
4	95%	79%	79%	84%
5	63%	95%	95%	100%
6	58%	100%	100%	100%
Overall Rating	42%	79%	84%	100%

[1] For school district A, the same procedures were used at all three grade levels, so agreement was the same. In school district B, different processes were used at different grade levels, so quality criteria ratings would not necessarily be the same across grades.

Table 2.

Percent of raters (n=19) providing comments for criteria in District A.

Criterion		
1	53%	A better description of the process and results
	32%	Provide evidence of sufficient coverage
	32%	No comments
2	63%	No comments
	32%	Additional documentation on the panel and process
3	47%	No comments
	32%	More information about the bias review panel and process
4	53%	More information about the panel and process
	32%	No comments
5	63%	Recommended that results be presented
	42%	Clarification of which two assessments were compared
	26%	Suggested additional strategies for measuring reliability
	21%	Clarification of how the scoring rubric was pre-tested
6	89%	Evidence that difficulty was considered in the process
	21%	Description of the rubric method

Table 3.

Percent of raters (n=19) providing comments for criteria across grades in District B.

Criterion		
1	68%	A better description of the process and results
	37%	Provide evidence of sufficient coverage
	21%	Suggested an independent panel for reviewing alignment
	21%	No comments
2	74%	Additional documentation on the panel and process
	21%	Provide results of the curriculum alignment
3	47%	More information about the bias review panel and process
	26%	No comments
4	63%	More information about the panel and process
	21%	Report the results of reviews and any decisions
	21%	No comments
5	74%	Commented on the lack of procedural documentation
	42%	Recommendation to present results of analyses
6	79%	Commented on the lack of any documentation